한국음향학회지 제41권 제2호 pp. 122~129 (2022)
The Journal of the Acoustical Society of Korea Vol.41, No.2 (2022)
https://doi.org/10.7776/ASK.2022.41.2.122

pISSN : 1225-4428
eISSN : 2287-3775

로컬 프레임 속도 변경에 의한 데이터 증강을 이용한 트랜스포머 기반 음성 인식 성능 향상

Improving transformer-based speech recognition performance using data augmentation by local frame rate changes

임성수,[1] 강병옥,[2] 권오욱[3†]

(Seong Su Lim,[1] Byung Ok Kang,[2] and Oh-Wook Kwon[3†])

[1]충북대학교 제어로봇공학전공, [2]한국전자통신연구원 인공지능연구소, [3]충북대학교 지능로봇공학과

(Received September 28, 2021; revised December 20, 2021; accepted December 30, 2021)

초 록: 본 논문은 프레임 속도를 국부적으로 조절하는 데이터 증강을 이용하여 트랜스포머 기반 음성 인식기의 성능을 개선하는 방법을 제안한다. 먼저, 원래의 음성데이터에서 증강할 부분의 시작 시간과 길이를 랜덤으로 선택한다. 그 다음, 선택된 부분의 프레임 속도는 선형보간법을 이용하여 새로운 프레임 속도로 변경된다. 월스트리트 저널 및 LibriSpeech 음성데이터를 이용한 실험결과, 수렴 시간은 베이스라인보다 오래 걸리지만, 인식 정확도는 대부분의 경우에 향상됨을 보여주었다. 성능을 더욱 향상시키기 위하여 변경 부분의 길이 및 속도 등 다양한 매개변수를 최적화하였다. 제안 방법은 월스트리트 저널 및 LibriSpeech 음성 데이터에서 베이스라인과 비교하여 각각 11.8 % 및 14.9 % 의 상대적 성능 향상을 보여주는 것으로 나타났다.

핵심용어: 종단간 음성 인식, 데이터 증강, 선형보간, 프레임 속도 변경

ABSTRACT: In this paper, we propose a method to improve the performance of Transformer-based speech recognizers using data augmentation that locally adjusts the frame rate. First, the start time and length of the part to be augmented in the original voice data are randomly selected. Then, the frame rate of the selected part is changed to a new frame rate by using linear interpolation. Experimental results using the Wall Street Journal and LibriSpeech speech databases showed that the convergence time took longer than the baseline, but the recognition accuracy was improved in most cases. In order to further improve the performance, various parameters such as the length and the speed of the selected parts were optimized. The proposed method was shown to achieve relative performance improvement of 11.8 % and 14.9 % compared with the baseline in the Wall Street Journal and LibriSpeech speech databases, respectively.

Keywords: End-to-end speech recognition, Data augmentation, Linear interpolation, Frame rate change

PACS numbers: 43.70.Dn, 43.72.Bs, 43.72.Ne

I. 서 론

최근 딥러닝이 다양한 분야에서 성능향상을 보이며 주목받게 됨에 따라, 음성 인식 분야에서도 딥러닝 기술을 사용하여 음성 신호로부터 직접 텍스트에 매칭시켜 학습하는 종단간 방식이 활발히 연구되고 있다. 딥러닝을 이용한 음성 인식의 경우, 기존 음성 인식보다 구조는 간단하지만 학습 데이터의 양에 의해 성능이 크게 좌우된다. 즉, 딥러닝은 많은 양의 레이블이 있는 학습 데이터가 필요하다. 하지만 이러한 대

†**Corresponding author:** Oh-Wook Kwon (owkwon@cbnu.ac.kr)
Department of Intelligent Systems and Robotics, Chungbuk National University, Chungdae-ro 1, Seowon-gu, Cheongju 28644, Republic of Korea
(Tel: 82-43-261-3374, Fax: 82-43-268-2386)

량의 레이블 데이터를 확보하기 위해서는 긴 시간의 작업이 필요하다. 이러한 문제로 인해, 최근에 더 많은 학습 데이터를 확보하기 위한 연구가 진행되고 있는데, 그 중에 한 접근 방식이 데이터 증강 방법이다.

데이터 증강 방법은 레이블 데이터에서 음성 부분을 변형시켜 새로운 레이블 데이터를 만드는 방법이다. 관련된 선행 연구로는 속도 섭동,[1] SpecAugment,[2] SpecSwap[3] 등이 있다. 그 중에서도 SpecAugment와 속도 섭동이 널리 사용되고 있다. SpecAugment는 음성 인식기의 입력인 스펙트로그램을 랜덤으로 마스크를 씌워 변형하는 방법이고, 속도 섭동은 기존 레이블 데이터를 리샘플링하여 원래 속도의 90 %와 110 %인 두 개의 추가 복사본을 만들어 총 3배 용량의 레이블 데이터(3-fold)를 학습에 이용하는 방법이다.

본 논문에서는 속도 섭동과 비슷한 방법으로 음성 데이터의 속도를 변경하여 새로운 전사 데이터를 만들어 전사 데이터양을 늘린다. 하지만 속도 섭동과 다르게 전체 발화의 속도를 변경하는 것이 아닌 랜덤하게 선택된 구간에서만 속도를 변경하게 되며, 파형 영역이 아닌 스펙트로그램 영역에서 속도를 변경한다. 속도를 변경하는 구간과 길이는 SpecAugment와 같이 랜덤하게 결정된다.

제안 방법은 특정 구간에서의 속도를 바꾸어 주기 때문에 발화의 전체 속도를 변경하여 길이를 변경해주는 속도 섭동에 비해 강인하고, 기존 스펙트로그램의 프레임을 변경하여 사용하기 때문에 추가적인 전처리 과정이 필요하지 않다는 장점이 있다. 또한 제안 방법은 SpecAugment와 다르게 마스크를 씌우지 않고 특정 구간의 속도를 변경하여 과소적합을 이루기 때문에 발화 길이에 대해 강인하도록 학습한다.

제안 방법의 성능을 평가하기 위하여, Wall Street Journal(WSJ),[4] LibriSpeech[5]의 음성 데이터베이스를 사용하였다. 제안 방법은 WSJ의 경우에 속도 섭동보다 Word Error Rate(WER)이 상대적으로 10.1 % 성능이 향상됨을 보여준다.

2장에서 음성인식 모델과 기존의 데이터 증강 방법에 대하여 설명하고, 3장에서는 제안된 데이터 증강 방법에 대하여 설명한다. 4장에서는 데이터베이스와 실험에 관해 설명하며, 5장에서 결론을 맺는다.

II. 선행 연구

본 장에서는 실험에 사용한 모델 구조와 기존의 데이터 증강 방법에 관하여 서술한다.

2.1 Transformer－CTC 음성 인식기

최근 딥러닝 음성 인식기의 구조는 한 발화 전체를 한 번에 받아 특징으로 만들어 주는 인코더와 음성 특징을 텍스트로 만드는 디코더로 구성된 인코더-디코더 구조가 좋은 성능을 보여준다. 인코더에서 Long Short-Term Memory(LSTM)을 사용하는 Listen, Attend and Spell(LAS)[6] 기반 음성 인식기가 먼저 제안되었지만 회귀구조를 가지기 때문에 병렬처리가 불가능하다. 이를 보완하기 위하여 인코더에서 행렬 곱을 사용하여 특징을 만드는 자기 주의 메커니즘을 이용하여 LAS에서 LSTM을 제거한 구조인 Transformer[7]가 제안되었다. LSTM을 제거함으로써 학습 시간을 단축하였으며, Transformer 인코더에 CTC[8]를 결합함으로써 정렬 문제를 완화하여 학습 속도를 더욱 단축시키고 정확도를 향상시킨다.[9]

2.2 속도 섭동[1]

입력으로 들어오는 음성 데이터의 속도를 변경시켜 데이터 증강을 하는 방법이다. Sox[10]를 사용하여 원래 속도의 90 %와 110 %로 리샘플링한다. 속도 섭동된 학습데이터는 기존의 학습데이터에 추가되어, 최종적으로 기존의 3배 용량인 학습데이터를 학습에 이용한다.

2.3 SpecAugment[2]

SpecAugment는 입력으로 들어오는 음성 데이터의 일부분을 랜덤으로 마스크를 씌워 왜곡한 데이터를 학습에 이용하는 데이터 증강 방법이다. 왜곡시키는 절차는 Fig. 1과 같이, 1)시간 워핑, 2)주파수 마스킹, 3)시간 마스킹으로 총 3가지 단계로 구성되어 있다.

시간 워핑은 한 발화에 대한 스펙트로그램의 프레임 개수를 L이라고 하면 사용자가 정의한 시간 워핑 파라미터 W에서 L-W까지의 범위에서 이미지 중심을 통과하는 임의의 한 점을 랜덤으로 선택한 후

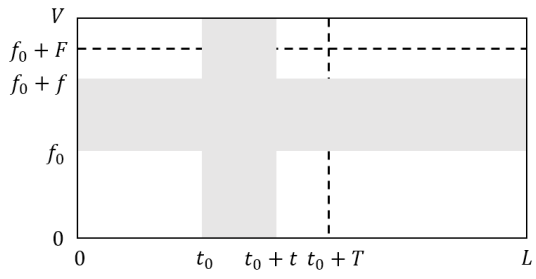

Fig. 1. SpecAugment applied to spectrogram. Spectrogram is time-masked between $[t_0,\ t+t_0)$ and frequency-masked between $[f_0,\ f_0+f)$.

왼쪽 또는 오른쪽으로 0에서 W까지의 범위에서 선택된 정수 w만큼 워핑하는 방법이다.[2]

주파수 마스킹은 정수 f만큼 연속적으로 주파수 마스킹을 하는 방법이다. 0에서부터 사용자가 정의한 주파수 마스크 파라미터 F까지의 범위에서 f를 선택하고 주파수 마지막 특징을 V라고 한다면 [0, V-f)에서 임의의 f_0을 선택한다. 선택한 f_0, f를 이용하여 [f_0, f_0+f) 범위에서 주파수 마스킹을 한다.

마지막으로 시간 마스킹은 주파수 마스킹과 비슷한 방법으로 다른 점은 주파수 축이 아니라 시간 축에 마스킹을 씌우는 방법이다. 즉, 정수 t만큼 연속적으로 시간 마스킹을 하는 방법이다. 0에서부터 사용자가 정의한 시간 마스크 파라미터 T까지의 범위 내에서 t를 선택하고 [0,L-t)에서 t_0를 선택한다. 선택한 t_0, t를 이용하여 [t_0, t+t_0) 범위에서 시간 마스킹을 한다.

SpecAugment 데이터 증강 방법으로 학습을 하게 될 경우 랜덤하게 입력 음성의 특정 부분이 변경되거나 마스킹 되어 학습에 이용되므로 과적합을 방지되는 효과가 존재하여 결과적으로 성능 향상이 이루어진다.

2.4 선형 보간법

보간은 주어진 값이 있을 때 그 사잇값을 추정하는 것이다. 선형 보간법은 보간하는 방법 중 하나로, 주어진 두 개의 값을 선형적으로 이어 사잇값을 구하는 방법이다. 두 값 (x_0,y_0)과 (x_1,y_1)이 주어졌을 때, 선형 보간법은 사잇값 x, y을 구하기 위해 다음과 같이 계산한다.

$$y = y_0 + (y_1 - y_0)\frac{x - x_0}{x_1 - x_0}. \quad (1)$$

III. 제안 방법

본 논문에서 제안하는 데이터 증강 방법은 다른 데이터 증강 방법과 마찬가지로 한정된 레이블 데이터를 변형시켜 기존보다 많은 학습 데이터를 확보하여 학습에 이용하는 것이 목표이다. 제안 방법은 음성 데이터 속도를 변경하여 학습에 이용하는 속도 섭동과 개념이 비슷하지만, 제안 방법의 경우 속도 섭동과 다른 방법으로 발화 길이를 변경하였다.

속도 섭동에서는 기존 레이블 데이터의 발화 전체 속도를 변경하여 새로운 학습 데이터로 이용하였다. 하지만 본 논문에서 제안 방법은 전처리 부분에서 발화 전체의 속도를 변형시키는 것이 아니라 스펙트로그램의 임의의 발화 구간에서 프레임 속도를 변경하여 학습에 이용한다. 따라서 제안 방법에서 속도 변경 정도 또는 구간 길이 가변을 통하여 데이터 증강 정도를 조절할 수 있다. 제안 방법은 반복 적용할 수 있으며, 적용될 때 중첩될 수 있다.

Fig. 2는 제안한 데이터 증강을 순서도로 나타낸 것이다.

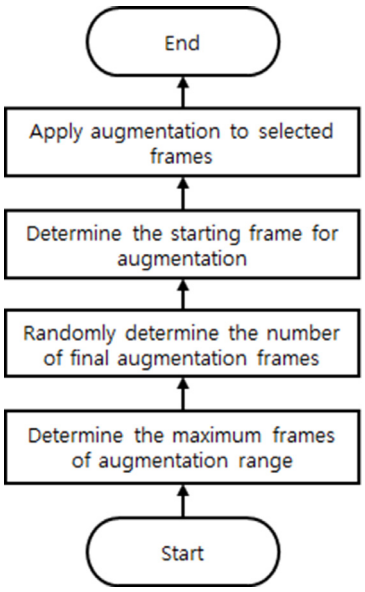

Fig. 2. Flow chart of the proposed data augmentation method.

3.1 알고리즘

제안 방법은 랜덤으로 선택된 값들을 사용하여 특정 구간의 프레임 속도를 변경하는 데이터 증강이다. 제안 방법은 발화의 길이에 따라 증강 정도를 비슷하게 적용하기 위해 발화 길이에 따라 증강 범위를 다르게 결정한다.

증강 적용 범위를 결정할 때 발화 길이 L에 비례하도록 최대 증강 프레임 개수 M을 다음과 같이 정의한다.

$$M = \lfloor mL \rfloor, \tag{2}$$

여기서 m은 전체 발화에서 증강하는 구간 비율을 나타내는 파라미터이다. 증강할 최종 프레임 개수 l은 $[0, M)$ 범위에서 다음과 같이 랜덤한 값으로 결정한다.

$$l \sim \mathrm{Uniform}(0, M), \tag{3}$$

여기서 $\mathrm{Uniform}(a,b)$는 $[a,b)$ 사이의 랜덤 정수를 반환하는 이산 균등 확률분포를 나타낸다. 증강 시작 위치 t_A는 증강되는 프레임 위치가 기존 프레임 위치를 넘어갈 수 없으므로 Eq. (3)의 샘플값을 이용하여 다음과 같이 랜덤한 값으로 결정한다.

$$t_A \sim \mathrm{Uniform}(0, L-l). \tag{4}$$

Fig. 3은 Eqs. (2)~(4)에서 정의한 파라미터를 원래 스펙트로그램에 나타낸 것이다.

Eqs. (3), (4)를 통해 결정한 t_A, l을 사용하여 $[t_A, t_A+l)$ 범위에서 본래 프레임 값을 대체 프레임으로 교체하여 프레임 속도를 바꾼다.

변경하려는 프레임 속도 S는 유리수로 다음과 같이 표현한다.

$$S = \frac{Q}{P}. \tag{5}$$

P와 Q는 각각 S의 분모와 분자를 나타낸다. 그리고 대체 프레임 위치 집합을 E라고 정의한다. 다음은 E를 구하는 방법을 나타낸다.

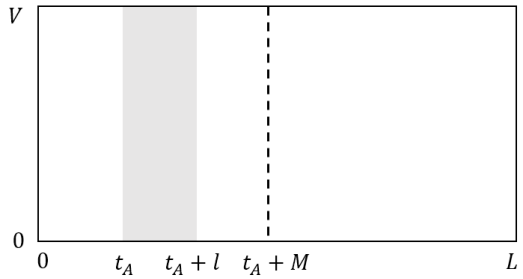

Fig. 3. Proposed method applied to the spectrogram when $m = 1/4$. The frame interval $[t_A, t_A+l)$ indicates the region where the frame rate is changed.

$$U = \frac{P}{Q}. \tag{6}$$

$$E = t_A + Uk, \quad k = 0,1,\cdots,N-1 \ \ (E < t_A + l). \tag{7}$$

예를 들어, $P=3$이고 $Q=2$이라면 $E = t_A$, t_A+1.5, …가 된다. E에서의 한 점을 e라고 할 때, e에서 소수 부분을 α로 정수 부분을 n으로 정의한다.

$$n = \lfloor e \rfloor. \tag{8}$$

$$\alpha = e - \lfloor e \rfloor. \tag{9}$$

대체 프레임의 위치가 e일 때, 대체 프레임 값 $f[e]$은 선형 보간법을 이용하여 다음과 같이 계산한다.

$$f[e] = (1-\alpha) f[n] + \alpha f[n+1]. \tag{10}$$

다차원으로 구성되어 있는 프레임에서 보간법을 사용할 때, 각각의 특징 차원의 값은 다른 특징 차원의 값에 영향을 주지 않는다. Fig. 4는 증강 프레임 구간에 있는 한 특징 차원에서 선형 보간법을 이용하여 대체 프레임 값을 구하는 방법을 보인다.

Fig. 5는 한 발화의 기존 스펙트로그램과 제안된 데이터 증강 방법이 적용된 후의 스펙트로그램을 비교한 것이다.

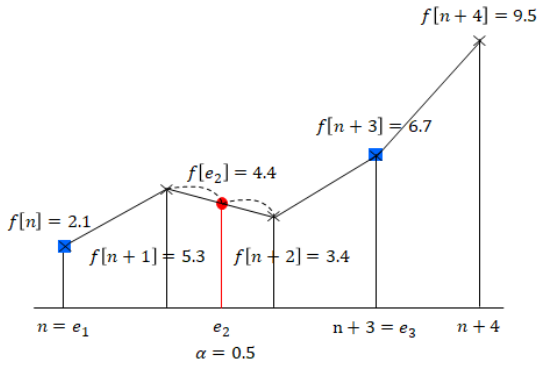

Fig. 4. (Color available online) Alternative frame values obtained through linear interpolation when $S=\frac{2}{3}$. The circles indicate alternative frame values, the crosses indicate the original frame values and the squares indicate the alternative frame values which are the same as the original values.

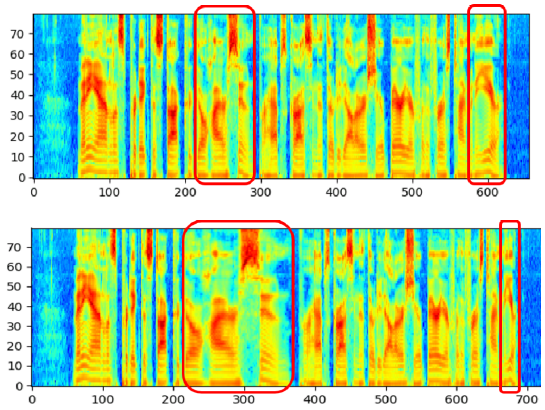

Fig. 5. (Color available online) Spectrogram of an utterance before (top) and after (bottom) applying the proposed method. The red boxes indicate the region where the frame rate was changed. Note that the augmented regions may overlap.

IV. 실험 결과

제안된 데이터 증강 방법에서 최적의 파라미터를 찾기 위한 실험과 제안된 방법이 성능 향상이 됨을 보이기 위하여 다양한 실험을 진행하였다.

4.1 데이터베이스

본 논문에서는 신뢰성 있는 결과를 위하여 총 2개의 데이터베이스를 사용하여 음성 인식 실험을 진행하였다.

WSJ은 영어 음성 데이터베이스로서, 약 81 h, 37,416개의 발화로 구성된 학습 데이터 세트“si284”로 학습을 진행하였고 약 1 h인 공식 검증 데이터 세트“dev93”과 약 0.7 h인 평가 데이터 세트“eval92”를 사용하여 성능을 평가하였다.

LibriSpeech는 영어 데이터베이스로서, 약 960 h로 구성되어 있지만 본 논문에서는 100 h로 학습을 진행하였다. 5.4 h로 구성되어 있는 LibriSpeech 검증 데이터 세트인 “dev-clean”과 5.4 h로 구성된 평가 데이터 세트 “test-clean”을 사용하여 성능을 평가하였다.

4.2 베이스라인

본 실험에서 사용한 음성 인식기는 ESPnet1 툴킷[11]이며, 모델 구조는 Transformer-CTC 구조를 사용하였다. Transformer는 각각 인코더 블록 12개, 디코더 블록 6개, 멀티-헤드 주의 집중은 8개로 설정하였다. 배치 크기는 64, 드롭아웃은 0.1, 결과 스무딩은 0.1을 사용하여 학습하였다. 성능을 도출한 모델은 100 이포크까지 학습을 한 후 검증 데이터 세트로 정확도를 계산하여 WSJ에서는 정확도가 높은 상위 10개의 모델을 LibriSpeech에서는 상위 5개의 모델을 평균하여 만든 모델을 사용하였다.

4.3 WSJ 데이터베이스 실험

본 절에서는 최적의 파라미터를 찾기 위한 실험들을 보여주고, 손실함수를 통해 제안 방법의 효과를 보여준다.

4.3.1 파라미터 변화에 따른 성능

3장에서 설명하였듯이, 제안 방법은 프레임 속도를 변경하여 데이터 증강을 한다. 본 실험에서는 파라미터 S에 따라 속도 조절을 하여 증강 정도를 변경하며 실험하였다.

Table 1은 프레임 속도에 따른 성능 평가 결과를 나타낸다. 속도뿐만 아니라 속도를 변경하는 구간의 길이에 따라 변동하는 속도에 따라 데이터 증강 정도가 변하게 된다. 프레임 속도를 늘리는 것과 줄이는 것에서 성능이 가장 좋은 파라미터는 각각 $S=2$, $S=1/2$로 나타났다.

Table 1. WER (%) of WSJ database according to different speed parameter S.

Speed	dev93	eval92
S = 1/2	**6.86**	**4.78**
S = 2/3	6.85	5.05
Baseline (S = 1)	7.88	5.02
S = 3/2	7.34	5.07
S = 2	7.54	4.93

Table 2. WER (%) of WSJ database according to different augment range m.

Speed	Ratio	dev93	eval92
S = 1/2	m = 1/1	7.32	5.21
	m = 1/2	**6.56**	**4.43**
	m = 1/4	6.86	4.78
	m = 1/8	6.80	4.94
S = 2	m = 1/1	7.52	5.46
	m = 1/2	7.27	4.70
	m = 1/4	7.54	4.93
	m = 1/8	7.15	4.87

Table 2는 구간의 길이에 따른 성능을 비교한 결과이다. 이전 실험에서 프레임 속도를 늘리는 것과 줄이는 것에서 성능이 가장 좋은 파라미터 S = 2, S = 1/2를 사용하여 데이터 증강 범위 교체 실험을 통해 구간의 길이에 따른 성능을 비교하였다. 변경 속도와 상관없이 m = 1/2 일 때, 즉 한 발화 길이의 반을 최대 증강 구간으로 잡았을 때의 성능이 가장 좋았다.

한 발화에 대해 제안 방법의 알고리즘을 여러 번 적용할 수 있다. 제안 알고리즘이 반복될 때는 증강 구간이 겹칠 수 있어 프레임 속도를 감소시키는 구간끼리 겹치거나 프레임 속도를 증가시키는 구간끼리 겹치게 되면 속도가 너무 빨라지거나 느려진다.

Table 3은 제안 방법의 반복 적용 회수에 따른 성능을 나타낸다. 본 실험에서 파라미터는 Table 2에서 성능이 좋은 m = 1/2로 설정하였으며, Table 3에 적힌 S를 순서에 상관없이 한 번씩 사용하여 제안 방법을 반복하였다. S = 1/2을 1회 적용했을 때의 결과가 제안 방법을 2회 반복했을 때의 결과보다 좋았으며, S = 1/2인 증강을 두 번 반복하였을 경우, 베이스라인보다도 상대적으로 성능이 약 12.9 % 나빠졌다.

Table 3. WER (%) of WSJ database according to different number of repetitions and speed parameter S.

Repetition	Speed (S)	dev93	eval92
1	1 (Baseline)	7.88	5.02
	1/2	**6.56**	**4.43**
2	1/2, 1/2	8.02	5.67
	2, 2	7.52	4.80
	2, 1/2	6.90	4.54

Table 4. WER (%) of WSJ database according to different data augmentation methods.

Data augmentation	dev93	eval92
No augmentation (Baseline)	7.88	5.02
SpecAugment	7.23	4.87
Speed perturb (3-fold)	7.24	4.93
Proposed method (S = 1/2)	**6.56**	**4.43**
Proposed method (S = 1/2, 2)	6.90	4.54

4.3.2 기존 데이터 증강과 비교

Table 4에서는 베이스라인 대비 기존 데이터 증강과 제안 방법의 성능 향상 정도를 비교를 하였다. eval92에서 속도 섭동의 경우 베이스라인 대비 WER이 상대적으로 약 1.8 %, SpecAugment는 3.0 %, S = 1/2을 적용한 제안 방법은 약 11.8 % 성능이 향상되었다. 결과적으로, 제안 방법은 속도 섭동에 비해서 상대적으로 약 10.1 % 성능이 향상되었고, SpecAugment를 적용하였을 때보다 약 9.0 % 성능이 향상되었다.

4.3.3 손실함수

데이터 증강을 통해 레이블 데이터를 변형하여 학습에 이용하는 것은 모델의 과적합을 방지한다. Fig. 6은 모델을 학습할 때 사용한 CTC-Transformer 결합 손실함수[9] 값을 나타낸다. 그림에서는 데이터 증강을 적용하지 않은 베이스라인, 기존 속도 조절 방법인 속도 섭동과 SpecAugment를 적용한 경우, 그리고 S = 1/2인 제안 방법의 손실 함수를 비교한 것으로서, 제안 방법의 손실 값이 다른 손실 값보다 더 천천히 그리고 더 작게 학습이 이루어지는 것을 보인다.

4.4 LibriSpeech 데이터베이스 실험

Table 5에서는 제안 방법 성능의 신뢰도를 높이기

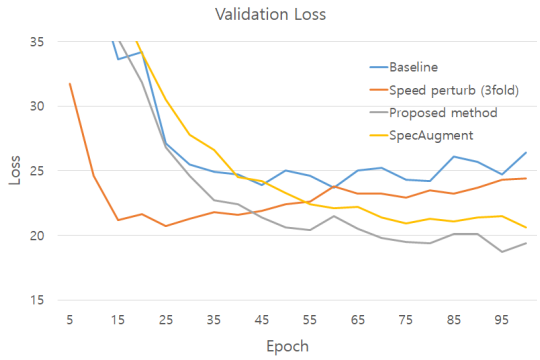

Fig. 6. (Color available online) Validation loss of baseline (blue), 3-fold Speed perturb (orange), SpecAugment (yellow) and proposed method (gray).

Table 5. WER (%) of 100 h LibriSpeech database according to different augmentation methods.

Data augmentation	dev-clean	test-clean
No augmentation (Baseline)	6.62	7.33
SpecAugment	6.23	6.96
Speed perturb (3-fold)	6.10	6.74
Proposed method (S = 1/2)	6.38	7.20
Proposed method (S = 1/2, 2)	**5.81**	**6.24**

위해 LibriSpeech 100시간의 데이터 세트를 사용하여 기존 데이터 증강과 제안 방법과 성능을 비교하였다. SpecAugment를 적용할 때 SpecAugment 파라미터는 각각 $W=5$, $F=30$, $T=40$, $n_mask=2$로 적용하였다.

기존 데이터 증강인 속도 섭동은 베이스라인과 비교하여 test-clean에서 WER이 상대적으로 약 8.0 % 성능이 향상되었고, SpecAugment는 약 5.0 % 성능이 향상되었다. 그에 비해 속도 $S=1/2, 2$를 적용한 제안 방법은 약 14.9 % 성능이 향상되었다. 결과적으로, 제안 방법은 속도 섭동에 비해서 상대적으로 약 7.4 % 성능이 향상되었고, SpecAugment에 비해 약 10.3 % 성능이 향상되었다.

V. 결 론

본 논문에서는 발화의 일부 구간에서 프레임 속도를 변경하는 새로운 데이터 증강 방법을 제안하였다. 발화의 길이가 다름에도 동일한 텍스트 결과가 나올 수 있다는 속도 섭동의 개념과 랜덤요소를 가진 SpecAugment의 개념을 동시에 가진 새로운 데이터 증강 방법이다. 제안 방법은 $S=1/2, 2$를 적용하였을 때, LibriSpeech 100 h, WSJ 데이터베이스 모두에서 기존 데이터 증강 방법인 SpecAugment와 속도 섭동보다 더 높은 성능 향상을 보여주었다. 향후에 선형 보간 이외에 다른 보간법을 통해 프레임 값을 계산하는 연구가 필요하다.

감사의 글

이 논문은 2019년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(2019-0-00004, 준지도학습형 언어지능 원천기술 및 이에 기반한 외국인 지원용 한국어 튜터링 서비스 개발).

References

1. T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, “Audio augmentation for speech recognition,” Proc. Interspeech, 3586-3589 (2015).
2. D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, “SpecAugment: A simple data augmentation method for automatic speech recognition,” arXiv:1904.08779 (2019).
3. X. Song, Z. Wu, Y. Huang, D. Su, and H. Meng, “SpecSwap: A simple data augmentation method for end-to-end speech recognition,” Proc. Interspeech, 581-585 (2020).
4. D. B. Paul and J. M. Baker, “The design for the wall street journal-based CSR corpus,” Proc. Speech and Natural Language Workshop, 357-362 (1992).
5. V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “LibriSpeech: An ASR corpus based on public domain audio books,” Proc. ICASSP. 5206-5210 (2015).
6. W. Chan, N. Jaitly, Q. V. Le, and O. Vinyals, “Listen, attend and spell,” arXiv:1508.01211 (2018).
7. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” Proc. NIPS. 5998-6008 (2017).
8. A. Graves, S. Fernandez, F. Gomez, and J. Schmidhuber, “Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks,” Proc. ICML. 369-376 (2006).
9. S. Kim, T. Hori, and S. Watanabe, “Joint CTC-attention

based end-to-end speech recognition using multi-task learning," Proc. ICASSP. 4835-4839 (2017).
10. *Sox, Audio Manipulation Tool*, http://sox.sourceforge.net/, (Last viewed March 25, 2015).
11. S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. E. Y. Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, "ESPnet: End-to-end speech processing toolkit," arXiv: 1804.00015 (2018).

저자 약력

▸ 임 성 수 (Seong Su Lim)

2020년 2월 : 충북대학교 전자공학부 학사
2022년 2월 : 충북대학교 제어로봇공학전공 석사
2022년 3월 ~ 현재 : ㈜퍼즐에이아이 연구원

▸ 강 병 옥 (Byung Ok Kang)

1997년 2월 : 포항공과대학교 전기전자공학과 학사
1999년 2월 : 포항공과대학교 전기전자공학과 석사
2017년 2월 : 충북대학교 전기·전자·컴퓨터학부 박사
1999년 ~ 2001년 : 삼성전자 무선사업부 선임연구원
2001년 ~ 현재 : 한국전자통신연구원 책임연구원

▸ 권 오 욱 (Oh-Wook Kwon)

1986년 2월 : 서울대학교 전자공학과 학사
1988년 2월 : 한국과학기술원 전기및전자공학과 석사
1997년 2월 : 한국과학기술원 전기및전자공학과 박사
1988년 3월 ~ 2000년 4월 : 한국전자통신연구원 책임연구원
2000년 5월 ~ 2001년 3월 : 한국과학기술원 연구교수
2001년 3월 ~ 2003년 8월 : UCSD 박사후 연구원
2003년 9월 ~ 현재 : 충북대학교 지능로봇공학과 교수